

Linear Combinations of Features as Leaf Nodes in Symbolic Regression

Supplementary Material

Jan Žegklitz and Petr Pošík

Introduction

This document is a guide to the supplementary material. In Sections 1 and 2 the datasets are described. In Sections 3 and 4 the full results on these problems are reported.

1 Toy problem datasets

1.1 S2D, S5D, S10D, RS2D, RS5D, RS10D

These are the toy problems on 2, 5 and 10 dimensions. The S* problems are a simple sigmoid function applied to the first variable, independent on the others. In the RS* problems, the sigmoid is rotated by $\frac{\pi}{4}$ in all pairs of axes, i.e. all variables become important. The problems are uniformly randomly sampled from the range $[-10, 10]^D$. There are $100 \cdot D$ samples in the training set and $250 \cdot D$ samples in the testing set.

2 Artificial and real-world datasets

2.1 K11C

Similar to Keijzer11 in [2] but with extra numerical coefficients throughout the formula:

$$f(\mathbf{x}) = (27.22x_1 - 4.54)(-0.39x_2) + 11.46 \sin((0.21x_1 - 1)(x_2 + 16.6) + 1.97).$$

The training set is 500 random samples drawn uniformly from the range $[-3, 3]^2$. The testing set is a grid in the same range with a spacing of 0.001 in each dimension (361 201 samples).

2.2 UB5D

Unwrapped Ball 5D [4] – a 5D artificial benchmark. The true relationship is

$$f(\mathbf{x}) = \frac{10}{5 + \sum_{i=1}^N (x_i - 3)^2}$$

where $N = 5$. The training set is 1024 random samples drawn uniformly from the range $[-0.25, 6.35]^5$. The testing set is 5000 samples obtained in the same way as the training set.

2.3 ASN

Airfoil Self-Noise, acquired from the UCI repository [1], is a 5D dataset regarding the sound pressure levels of airfoils based on measurements from a wind tunnel. Training/testing set comes from a random 0.7/0.3 split of the original dataset (1503 datapoints in total).

2.4 CCS

Concrete Compressive Strength [5], acquired from the UCI repository [1], is an 8D dataset representing a highly non-linear function of concrete age age and ingredients. Training/testing set comes from a random 0.7/0.3 split of the original dataset (1030 datapoints in total).

2.5 ENC, ENH

Energy Efficiency of Cooling/Heating [3], acquired from the UCI repository [1], are 8D datasets regarding the energy efficiency of cooling and heating of buildings. Training/testing set comes from a random 0.7/0.3 split of the original dataset (768 datapoints in total).

2.6 SU, SU-I

These two datasets come from the domain of reinforcement learning and represent the value functions of an inverted pendulum swing-up task, computed by a numeric approximator. Both datasets are 2D (pendulum angle and angular velocity) and the value (that is the regression target) is the value of the state w.r.t. the goal state which, for the SU variant is located at $[-\pi, 0]$ and equivalently $[\pi, 0]$ (due to the circular nature of the problem). The SU-I variant represents identical function but the angle coordinate is shifted by $\frac{\pi}{2}$. Training/testing set comes from a random 0.7/0.3 split of the original dataset (441 samples in total).

The plots of both datasets are in Figures 1 and 2. The full raw data (i.e. before splitting to training/testing sets) of both datasets are stored in the files `swingup.txt` (SU) and `swingup-inverted.txt` (SU-I). Each line of the file is one data point, the field delimiter is a tab character (ASCII 0x09), lines are terminated with CRLF (i.e. Windows style, ASCII 0x0D 0x0A). First two

columns are the angle and angular velocity, the last column is the value (i.e. the target value for the regression).

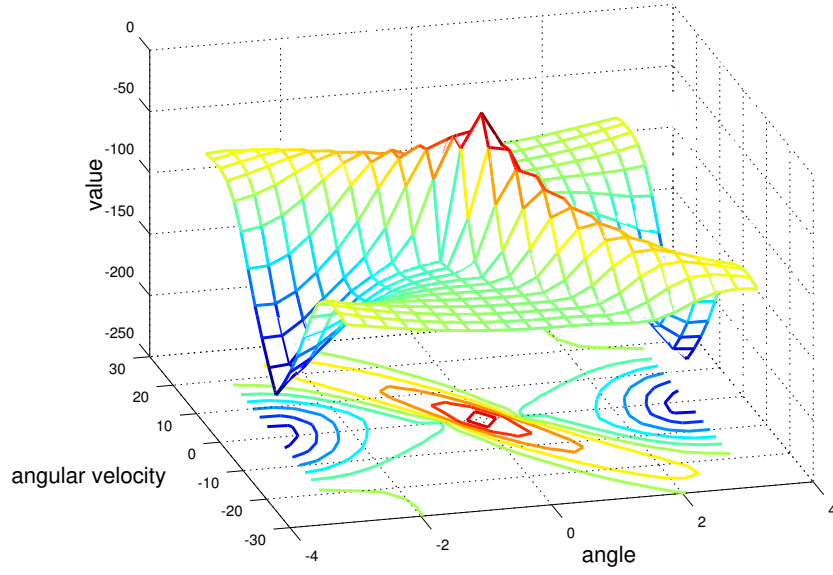


Figure 1: SU dataset plot.

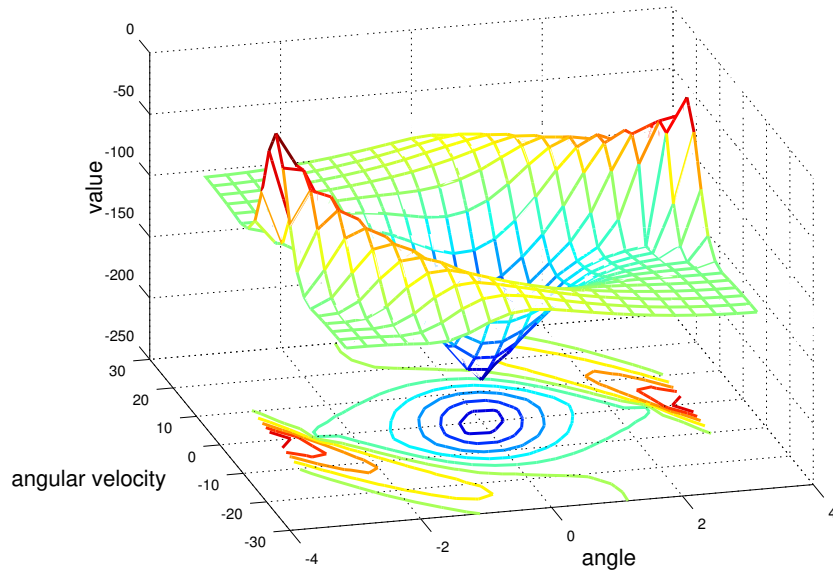


Figure 2: SU-I dataset plot.

2.7 MM

This dataset comes from the domain of reinforcement learning and represents the value function of a 2-coil magnetic manipulation task (a ball is manipulated by two electromagnetic coils in a linear space to a desired position), computed by a numeric approximator. It has 2 dimensions (the ball's position and velocity) and the value (that is the regression target) is the value of the state w.r.t. the goal state. Training/testing set comes from a random 0.7/0.3 split of the original dataset (729 samples in total).

The plot of the data is in Figure 3. The full raw data is stored in the file `magman.txt`. The format is identical to SU. The first two columns are the ball position and velocity, the last column is the value (i.e. the target value for the regression).

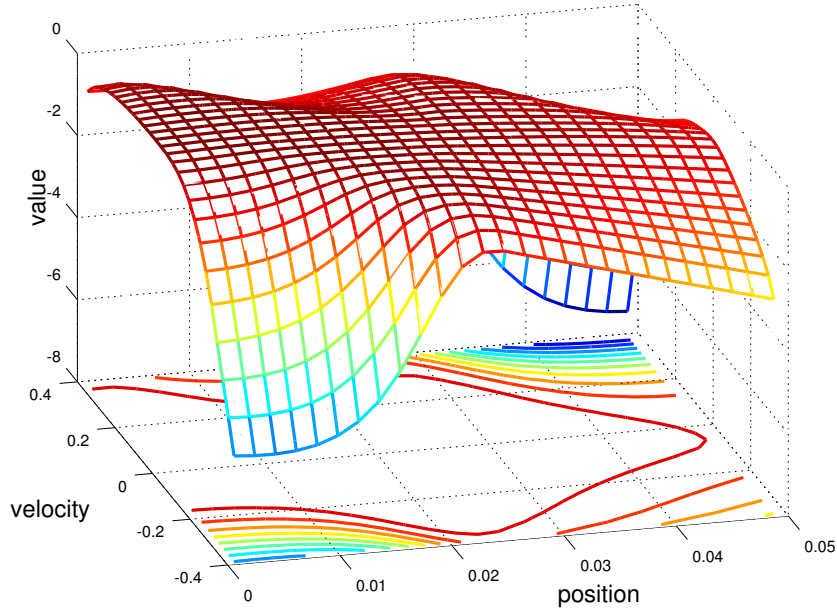


Figure 3: MM dataset plot.

3 Full results on toy problems

In Tables 1 through 6 contain results for the toy problems, including the 2D and 10D cases.

Table 1: Results on the S2D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	$\frac{\max}{\min}$	median	$\frac{\max}{\min}$		LCF	depth
- -	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0	2.73
UM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.509	2.57
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.503	3.4
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.524	3.37
SM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.509	2.57
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.47	4.33
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.411	4
GB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.385	4.6
GC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.342	4.8

Table 2: Results on the RS2D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	$\frac{\max}{\min}$	median	$\frac{\max}{\min}$		LCF	depth
- -	1	$\frac{1}{0.999}$	1	$\frac{1}{0.96}$		0	10.9
UM	1	$\frac{1}{0.995}$	1	$\frac{1}{0.987}$		0.499	10.7
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.931	4.37
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.919	4.27
SM	1	$\frac{1}{0.992}$	1	$\frac{1}{0.988}$		0.679	10.8
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.898	8.37
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.886	8
GB	1	$\frac{1}{0.963}$	1	$\frac{1}{0.95}$	✗	0.486	10
GC	0.999	$\frac{1}{0.965}$	0.999	$\frac{1}{0.907}$	✗	0.305	9.57

Table 3: Results on the S5D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	$\frac{\max}{\min}$	median	$\frac{\max}{\min}$		LCF	depth
- -	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0	4.33
UM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.482	4.1
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.576	3.5
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.521	3.77
SM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.469	4.13
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.475	4.2
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.48	3.73
GB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.38	4.23
GC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.352	4.87

Table 4: Results on the RS5D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	$\frac{\max}{\min}$	median	$\frac{\max}{\min}$		LCF	depth
- -	0.995	$\frac{0.997}{0.942}$	0.991	$\frac{0.996}{0.912}$		0	10.9
UM	0.993	$\frac{1}{0.93}$	0.99	$\frac{1}{0.907}$		0.524	10.6
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.98	4.6
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.974	4.63
SM	0.995	$\frac{1}{0.158}$	0.993	$\frac{0.999}{-0.168}$		0.579	10.9
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.9	8.1
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.954	7.77
GB	1	$\frac{1}{0.855}$	1	$\frac{1}{0.659}$	✓	0.817	8.53
GC	0.974	$\frac{1}{0.872}$	0.962	$\frac{1}{0.736}$	✗	0.651	6.7

Table 5: Results on the S10D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	$\frac{\max}{\min}$	median	$\frac{\max}{\min}$		LCF	depth
- -	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0	4.37
UM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.489	3.5
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.663	2.97
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.671	3.17
SM	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.465	3.87
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.469	3.6
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.54	3.87
GB	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.58	4.2
GC	1	$\frac{1}{1}$	1	$\frac{1}{1}$		0.341	5.13

Table 6: Results on the RS10D toy problem. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		vb	mean	mean
	median	\max_{\min}	median	\max_{\min}		LCF	depth
- -	0.984	$\frac{0.995}{0.838}$	0.98	$\frac{0.992}{0.812}$		0	10.9
UM	0.94	$\frac{0.993}{0.808}$	0.918	$\frac{0.99}{0.803}$		0.588	10.8
UB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.985	4.57
UC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.982	4.23
SM	0.967	$\frac{0.994}{0.549}$	0.96	$\frac{0.993}{0.512}$		0.517	10.8
SB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.938	6.7
SC	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.942	6.93
GB	1	$\frac{1}{1}$	1	$\frac{1}{1}$	✓	0.986	8.27
GC	0.991	$\frac{1}{0.915}$	0.99	$\frac{1}{-1.21e+17}$	✓	0.681	4.6

4 Full results on the realistic problems

Tables 7 through 15 contain the results for realistic problems including configurations UM, SM GB and GC.

Table 7: Performance on the K11C dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.981	$\frac{0.997}{0.971}$	0.976	$\frac{0.995}{0.965}$		0
UM	0.986	$\frac{0.997}{0.975}$	0.981	$\frac{0.996}{0.969}$	✓	0.542
UB	0.998	$\frac{0.999}{0.99}$	0.996	$\frac{0.999}{0.978}$	✓	0.873
UC	0.998	$\frac{1}{0.992}$	0.997	$-3.24e+29$	✓	0.874
SM	0.986	$\frac{0.997}{0.973}$	0.981	$\frac{0.997}{0.966}$	✓	0.595
S B	0.991	$\frac{0.998}{0.954}$	0.989	$\frac{0.998}{0.945}$	✓	0.603
S C	0.992	$\frac{0.998}{0.954}$	0.99	$\frac{0.997}{0.948}$	✓	0.622
GB	0.972	$\frac{0.993}{0.958}$	0.967	$\frac{0.992}{0.952}$		0.549
GC	0.971	$\frac{0.989}{0.956}$	0.966	$\frac{0.987}{0.949}$	✗	0.177

References

- [1] K. Bache and M. Lichman. *UCI Machine Learning Repository*. <http://archive.ics.uci.edu/ml>. 2013.
- [2] James McDermott et al. “Genetic Programming Needs Better Benchmarks”. In: *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*. GECCO ’12. Philadelphia, Pennsylvania, USA: ACM, 2012, pp. 791–798. ISBN: 978-1-4503-1177-9. DOI: 10.1145/2330163.2330273. URL: <http://doi.acm.org/10.1145/2330163.2330273>.
- [3] Athanasios Tsanas and Angeliki Xifara. “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools”. In: *Energy and Buildings* 49 (2012), pp. 560–567. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2012.03.003. URL: <http://dx.doi.org/10.1016/j.enbuild.2012.03.003>.

Table 8: Performance on the UB5D dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

	mode tuning	training R^2		testing R^2		vb	mean LCF
		median	max min	median	max min		
-	-	0.885	0.976 0.808	0.866	0.968 0.796		0
UM		0.884	0.965 0.821	0.862	0.966 0.802		0.539
UB		0.857	0.887 0.828	0.828	0.856 0.58	X	0.823
UC		0.858	0.932 0.824	0.826	0.892 0.807	X	0.802
SM		0.907	0.982 0.813	0.89	0.977 0.805		0.533
S B		0.839	0.972 0.802	0.816	0.967 0.796	X	0.553
S C		0.839	0.93 0.816	0.818	0.908 0.795	X	0.601
GB		0.825	0.881 0.761	0.808	0.873 0.749	X	0.334
GC		0.828	0.902 0.778	0.808	0.887 0.783	X	0.0683

- [4] E.J. Vladislavleva, G.F. Smits, and D. den Hertog. “Order of Nonlinearity as a Complexity Measure for Models Generated by Symbolic Regression via Pareto Genetic Programming”. In: *Evolutionary Computation, IEEE Transactions on* 13.2 (Apr. 2009), pp. 333–349. ISSN: 1089-778X. DOI: 10.1109/TEVC.2008.926486.
- [5] I.-C. Yeh. “Modeling of strength of high-performance concrete using artificial neural networks”. In: *Cement and Concrete Research* 28.12 (1998), pp. 1797–1808. ISSN: 0008-8846. DOI: 10.1016/S0008-8846(98)00165-3. URL: [http://dx.doi.org/10.1016/S0008-8846\(98\)00165-3](http://dx.doi.org/10.1016/S0008-8846(98)00165-3).

Table 9: Performance on the ASN dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.842	0.892 0.72	0.824	0.885 0.625		0
UM	0.845	0.89 0.787	0.824	0.89 0.729		0.461
UB	0.849	0.914 0.729	0.818	0.893 -0.719		0.834
UC	0.841	0.894 0.705	0.818	0.88 0.623		0.828
SM	0.836	0.887 -7.23	0.811	0.884 -4.14		0.462
S B	0.804	0.842 0.675	0.77	0.829 0.624	X	0.651
S C	0.8	0.867 0.71	0.76	0.861 0.653	X	0.68
GB	0.817	0.86 0.631	0.788	0.88 0.584		0.43
GC	0.778	0.849 0.669	0.757	0.859 0.645	X	0.309

Table 10: Performance on the CCS dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.869	0.89 0.848	0.844	0.868 $-8.68e+07$		0
UM	0.866	0.885 0.851	0.839	0.865 $-1.16e+06$		0.496
UB	0.901	0.924 0.869	0.859	0.892 0.806	✓	0.87
UC	0.899	0.931 0.854	0.858	0.88 0.758	✓	0.885
SM	0.862	0.885 0.846	0.837	0.882 0.799		0.467
S B	0.889	0.906 0.868	0.851	0.898 $-4.74e+04$		0.676
S C	0.893	0.908 0.857	0.846	0.873 -291		0.707
GB	0.859	0.885 0.844	0.825	0.867 -99.6		0.43
GC	0.854	0.867 0.836	0.83	0.879 $-7.35e+06$		0.252

Table 11: Performance on the ENC dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.974	0.981 0.969	0.97	0.982 0.963		0
UM	0.974	0.984 0.971	0.97	0.98 0.961		0.548
UB	0.974	0.988 0.97	0.969	0.982 0.957		0.751
UC	0.975	0.986 0.972	0.971	0.985 0.961		0.772
SM	0.974	0.985 0.97	0.969	0.979 0.96		0.52
SB	0.974	0.979 0.969	0.968	0.973 0.965		0.609
SC	0.973	0.98 0.969	0.968	0.976 0.962		0.609
GB	0.971	0.98 0.967	0.967	0.981 0.954	X	0.551
GC	0.971	0.973 0.968	0.967	0.972 0.957		0.123

Table 12: Performance on the ENH dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.998	0.998 0.996	0.997	0.998 0.995		0
UM	0.998	0.998 0.996	0.997	0.998 0.995		0.501
UB	0.997	0.998 0.993	0.997	0.998 0.991		0.73
UC	0.998	0.998 0.995	0.997	0.998 0.994		0.732
SM	0.997	0.998 0.996	0.997	0.998 0.996		0.546
SB	0.997	0.998 0.993	0.997	0.998 0.993		0.592
SC	0.997	0.998 0.99	0.997	0.998 0.988		0.61
GB	0.996	0.998 0.99	0.996	0.997 0.99	X	0.487
GC	0.996	0.998 0.989	0.996	0.997 0.986	X	0.0995

Table 13: Performance on the SU dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.955	0.988 0.879	0.909	0.978 -0.664		0
UM	0.96	0.993 0.877	0.918	0.987 0.683		0.527
UB	0.985	0.994 0.963	0.971	0.994 0.881	✓	0.894
UC	0.985	0.996 0.93	0.966	0.992 0.916	✓	0.885
SM	0.946	0.987 0.846	0.907	0.981 -0.185		0.528
S B	0.977	0.991 0.881	0.955	0.984 0.819	✓	0.598
S C	0.968	0.993 0.885	0.958	0.978 0.694		0.633
GB	0.927	0.983 0.727	0.885	0.985 -5.11e+03		0.466
GC	0.88	0.961 0.805	0.822	0.941 0.627	✗	0.165

Table 14: Performance on the SU-I dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.931	0.979 0.841	0.885	0.97 0.175		0
UM	0.938	0.992 0.886	0.899	0.975 0.606		0.517
UB	0.97	0.993 0.938	0.955	0.987 0.886	✓	0.895
UC	0.976	0.991 0.915	0.962	0.988 0.865	✓	0.912
SM	0.937	0.989 0.851	0.91	0.98 -6.84		0.498
S B	0.942	0.988 0.884	0.928	0.992 0.769		0.569
S C	0.952	0.989 0.836	0.931	0.99 0.788	✓	0.623
GB	0.887	0.966 0.694	0.853	0.967 0.505		0.518
GC	0.862	0.925 0.768	0.829	0.915 -2.86e+03	✗	0.151

Table 15: Performance on the MM dataset. Column titled “vb” stands for “versus baseline” and signifies whether the result is statistically significantly better than the baseline. Column titled “mean LCF” shows mean fraction of non-constant leaf nodes that are LCFs.

mode tuning	training R^2		testing R^2		mean	
	median	max min	median	max min	vb	LCF
- -	0.966	0.987 0.954	0.96	0.983 0.93		0
UM	0.97	0.988 0.957	0.961	0.987 0.947		0.552
UB	0.988	0.997 0.973	0.985	0.995 0.969	✓	0.763
UC	0.988	0.996 0.969	0.985	0.995 0.943	✓	0.797
SM	0.97	0.989 0.958	0.965	0.982 0.937		0.565
S B	0.976	0.991 0.967	0.973	0.986 0.961	✓	0.559
S C	0.974	0.997 0.947	0.971	0.996 0.935	✓	0.563
GB	0.967	0.982 0.941	0.959	0.981 0.928		0.462
GC	0.961	0.977 0.942	0.952	0.969 0.928		0.248